GHAJAR
EXHIBIT 29

Page 1

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

_____

RICHARD KADREY, et al.,   )

        Individual and   )

        Representative   )

        Plaintiffs,   )

)

  v.   )   Case No.:

)   3:23-cv-03417-VC

META PLATFORMS, INC.,   )

        Defendant.   )

_____)

** HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY **

Videotaped 30(b)(1) deposition of

NIKOLAY BASHLYKOV

Volume I

Thursday, December 5, 2024

London, England

United Kingdom

Reported stenographically by:

Leah M. Willersdorf,

RMR, CRR, FBIVR, ACR, QRR2*, CLR

_____

DIGITAL EVIDENCE GROUP

1730 M. Street, NW, Suite 812

Washington, D.C. 20036

(202) 232-0646

Page 12

```
 1   working on, like, language modeling.

 2        Q.    Is that the GenAI group?

 3        A.    Correct.

 4        Q.    GenAI group within Meta?

 5        A.    Correct.

 6        Q.    So would it be fair to say that you are

 7   also employed by Meta?

 8        A.    That legal point, I -- yeah, I don't know.

 9        Q.    Okay.

10             MR. WEINSTEIN:  I think it's easier just

11   to refer to his employer as Meta rather than Facebook

12   UK, so...

13             MR. YOUNG:  Okay.

14             THE WITNESS:  Yeah.

15             MR. YOUNG:  Great.

16   BY MR. YOUNG:

17        Q.    And you testified that you are currently a

18   software engineer at Meta; is that correct?

19        A.    Correct.

20        Q.    Okay.  And how long have you been in that

21   role?

22        A.    Since August twenty twenty -- I just don't
```

12/5/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Nikolay Bashlykov, Vol I
Highly Confidential - Attorneys' Eyes Only

Page 31

 1   correct?

 2   A.   Yes.

 3   Q.   Okay, thank you.

 4   So, sir, just looking at this paragraph

 5   here, it says you were a key contributor to LLaMA 2,

 6   LLaMA 3 and LLaMA 3.1, right?

 7   A.   Yes.

 8   Q.   Okay. So the LLaMA models, those are

 9   large language models that have been developed and

10   distributed by Meta, correct?

11   A.   Correct.

12   Q.   Okay. And the LLaMA 2 models -- the

13   LLaMA 2 series of models was released sometime in July

14   of 2023, correct?

15   A.   Yes.

16   Q.   Okay. And the LLaMA 3 series of models

17   was released sometime in January of this year,

18   correct?

19   A.   I don't remember the exact month.

20   Q.   Okay. But it was released sometime early

21   this year, correct?

22   A.   Correct.

Page 45

1   As part of your responsibilities, this is

2   with respect to LLaMA 3, was it also part of your job

3   responsibilities to write scripts for LLaMA 3?

4   A.   What do you mean "scripts by LLaMA 3"?

5   Q.   As part of your responsibility for

6   LLaMA 3, did you write source code for LLaMA 3?

7   A.   For some parts, I did.

8   Q.   Okay.   Now, did you also write source code

9   or scripts that removed, for example, PII in datasets?

10   A.   I had scripts for cleaning, yes.

11   Q.   Okay.   And by PII, you understand I mean

12   to be personally identifying information, for example

13   email addresses or phone numbers?

14   A.   Some of this information.

15   Q.   Okay.   Now, you also wrote scripts to

16   detect for repetition, correct?

17   A.   In our understanding of "repetition," yes.

18   Q.   Okay, thank you.

19   Now, you also wrote scripts to detect

20   copyright -- and by "copyright," I mean what Meta

21   understood as copyright -- correct?

22   MR. WEINSTEIN: Object to form.

Page 46

1   THE WITNESS: Could you repeat again

2   the question, please?

3   BY MR. YOUNG:

4   Q.   Yes.   Did you -- as part of your

5   responsibilities at LLaMA [sic], did you, sir, write

6   source code or scripts that you specifically called

7   copyright?

8   A.   The scripts were not called copyright.

9   Q.   Okay.   Were the scripts designed to detect

10   copyright material -- excuse me.

11   Were the scripts or source code that you

12   wrote designed to detect words that commonly would

13   indicate copyright and copyrighted material?

14   A.   The scripts, as far as I recollect,

15   identified "copyright" as a word and probably a few

16   other symbols.

17   Q.   Okay.   Like, for example, the C symbol

18   within the circle, was that one of the icons that you

19   identified?

20   A.   That could be, if I see the source code,

21   yes.

22   Q.   Okay.   And this script that we're talking

12/5/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Nikolay Bashlykov, Vol I
Highly Confidential - Attorneys' Eyes Only

Page 47

1   about, how it generally worked was that it would

2   search data for specific words and if it detected that

3   word, it would remove the line of the data that it

4   detected the word in, correct?

5   A.   It depends on the logic.   Either -- so the

6   logic is a bit more sophisticated than that.

7   Q.   Okay.   But forgiving my layperson's

8   understanding and description of the source code, what

9   I described is generally at a high level correct,

10   right?

11   A.   Yes.

12   Q.   Okay.   And if -- there were certain terms

13   that you would identify or that the script would

14   identify for flagging and removal, right?

15   A.   Sorry, could you repeat again?

16   Q.   Yes.   So there were certain terms that

17   the script would identify and remove the line if it

18   detected the word in it, right?

19   A.   Yes.

20   Q.   Okay.   And "copyright" being one of those

21   words, right?

22   A.   Yes.

Case 3:23-cv-03417-VC   Document 490-28   Filed 03/24/25   Page 8 of 13

12/5/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Nikolay Bashlykov, Vol I
                        Highly Confidential - Attorneys' Eyes Only

Page 48

1   Q.   Okay.   And certain symbols, for example,

2   as we discussed, right?

3   A.   Yes.

4   Q.   Okay.   In other words, that may be

5   commonly found in the beginning or the end of certain

6   books, correct?

7   MR. WEINSTEIN: Object to form.

8   THE WITNESS: In the script, there was --

9   okay.   Could you repeat the question?

10   BY MR. YOUNG:

11   Q.   Yes, absolutely.   Let me get a little more

12   specific.

13   So in the scripts that you wrote to detect

14   for words associated with copyright, or what Meta --

15   what you labeled as copyright, did it search a

16   particular portion in the front of the book and also

17   in the back of the book?

18   A.   As far as I remember, there was some logic

19   of checking specifically in the beginning and in the

20   end of the book.

21   Q.   Because in the beginning of the book is

22   typically where you see words like "ISBN" or "Table of

Page 49

1   Contents" or "Acknowledgments," correct?

2   A.   Usually in the beginning and in the end of

3   the book there are certain headers which are

4   unrelevant for training.

5   Q.   And this information would be information

6   such as the copyright date, right?

7   A.   That's part of it, yes.

8   Q.   Okay.   Or perhaps the name of the author?

9   A.   No.   We didn't check for the name of the

10   author.

11   Q.   But you did check for the word "author,"

12   right?

13   A.   I do not remember, but if you have the

14   source code, you can check.

15   Q.   Okay.   And you also checked for the word

16   "content," right?

17   A.   Again, I can only speculate because

18   I don't remember the code that I wrote, like, almost

19   two years ago, but we could have looked into the Table

20   of Contents.

21   Q.   Okay, thank you.   Okay.

22   Now, sir, you under -- you understand,

Case 3:23-cv-03417-VC   Document 490-28   Filed 03/24/25   Page 10 of 13

12/5/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Nikolay Bashlykov, Vol I
                          Highly Confidential - Attorneys' Eyes Only

Page 156

```
 1          A.   I think that is some general description.

 2    It might not be 100 percent correct, so it's better to

 3    refer to the script itself, if you...

 4          Q.   Okay.  But this is just you reporting

 5    generally, at a high level, what this script does?

 6          A.   Yes.  Somewhat, yes.

 7          Q.   Okay.  And then similarly, for PII data,

 8    right, it says "Remove emails," right?

 9          A.   Yes.

10          Q.   Okay.  So the PII script is just detecting

11    and removing email addresses from the data, correct?

12          A.   As far as I can read from here, and also

13    better to consult to the code base itself, it is only

14    this, yes.

15          Q.   Okay, great.  And then the next row talks

16    about copyright, right?

17          A.   Yes.

18          Q.   Okay.  Can you read for me that bullet

19    point, please.

20          A.   "Remove rows containing copyright in the

21    first and last 25% of the book."

22          Q.   Okay, great.  And then the next bullet
```

12/5/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Nikolay Bashlykov, Vol I
Highly Confidential - Attorneys' Eyes Only

Page 157

1   describes the words that you were actually searching

2   for -- that the script was actually searching for to

3   detect the rows to remove, correct?

4   A.   Yes.

5   Q.   Okay.   Can you read for me that bullet

6   point, please.

7   A.   "Rows containing any of these words," and

8   then in brackets 'ISBN,' 'Copyright,' symbol C, 'All

9   Rights Reserved,' 'DOI.'"

10   Q.   Okay.   Do you remember who picked these

11   words for the script?

12   A.   I think it was myself after manual

13   examination of the datasets.

14   Q.   Okay.   Did you choose these words in

15   consultation with anyone?

16   A.   No.

17   Q.   Okay.   Why did you pick these words to

18   detect and remove from the dataset?

19   A.   They commonly occurred in the books.

20   They do not, to my understanding, bring any value to

21   training and -- yeah.

22   Q.   Okay.   Why is it a problem if a word --

Page 158

1   if these words commonly appear, and why did you feel

2   like they needed to be removed from the dataset?

3   A.   So commonly appear is not a problem.

4   A problem is that, like, it's, rather, the -- like,

5   when you do filter, you look for particular patterns

6   in the documents and you also try to -- because you

7   can't filter all the useless tokens for training from

8   all the documents, so you're usually focusing on some

9   patterns that are prevalent in many of the documents,

10   because with one script you can easily remove a lot of

11   those unused token -- like, not-used-for-training

12   tokens.

13   So that was the main kind of idea behind

14   looking into some of the most commonly words that

15   present in the documents.

16   Q.   Okay, thank you.

17   Now, did you apply this copyright script

18   to any other dataset?

19   A.   I did not.

20   Q.   Okay. Do you know if others applied this

21   copyright script to other datasets?

22   A.   I do not know of such cases.

Page 244

1   CERTIFICATE OF COURT REPORTER

2   I, LEAH M. WILLERSDORF, Registered Merit

3   Reporter, Certified Realtime Reporter, Fellow of the

4   British Institute of Verbatim Reporters, Qualified

5   Realtime Reporter Level 2, and Certified LiveNote

6   Reporter, do hereby certify that:

7   NIKOLAY BASHLYKOV appeared before me on

8   Thursday, December 05, 2024, was sworn by me and was

9   thereupon examined by counsel; that the testimony of

10   said witness was taken by me stenographically; that

11   the foregoing is a true and accurate record to the

12   best of my knowledge, skill and ability; that I am

13   neither a relative nor employee of any party to the

14   action in which this testimony was taken; that I am

15   neither relative nor employee of any attorney or

16   counsel employed by any party thereto; and, further,

17   I am not financially or otherwise interested in the

18   outcome of the action.

19

20

21   LEAH M. WILLERSDORF

                 RMR, CRR, FBIVR, ACR, QRR2, CLR

22   (December 08, 2024)